

# Projected gradient descent algorithms for quantum state tomography

Eliot Bolduc<sup>1</sup>, George Knee<sup>2</sup>, Erik Gauger<sup>1</sup>, Jonathan Leach<sup>1\*</sup>

<sup>1</sup> *SUPA, Institute of Photonics and Quantum Sciences, Heriot-Watt University,  
David Brewster Building, Edinburgh, EH14 4AS, UK and*

<sup>2</sup> *Department of Physics, University of Warwick, Coventry, CV4 7AL, UK*

(Dated: January 2, 2017)

Accurate quantum tomography is a vital tool in both fundamental and applied quantum science. It is a task that involves processing a noisy measurement record in order to construct a reliable estimate of an unknown quantum state, and is central to quantum computing, metrology, and communication. To date, many different approaches to quantum state estimation have been developed, yet no one method fits all applications, and all fail relatively quickly as the dimensionality of the unknown state grows. In this work, we suggest that projected gradient descent is a method that can evade some of these shortcomings. We present three novel tomography algorithms that use projected gradient descent and compare their performance with state-of-the-art alternatives, i.e. the diluted iterative algorithm and convex programming. Our results find in favour of the general class of projected gradient descent methods due to their speed, applicability to large states, and the range of conditions in which they perform.

## INTRODUCTION

The reconstruction of an unknown quantum state – known as quantum tomography – is a fundamental task in quantum information science, where a myriad of new technologies promise to exploit special features of quantum states in order to enhance communication, metrology, and computation. Since the quantum state represents maximal information about a physical system, all physical properties can be calculated from it. Checking for the existence for a highly entangled state, a state which can violate a Bell inequality, or even the initial state required in a gate-based quantum computer are thus just some examples of the importance of inferring the quantum state from laboratory data. As experimental methods progress, the complexity of quantum systems that can be well controlled in the laboratory grows. In recent times, for example, various groups have been able to demonstrate quantum control of a high number of qubits [1–3]. To gain an idea of the challenge of state reconstruction, one need only consider that the number of real parameters required to describe the joint state of  $n$  qubits scales as order  $2^{2n}$ . Alternatively, the orbital angular momentum of single photons, for example, is a single degree of freedom with a large amount of internal structure: it has recently been characterised via the reconstruction of a 100,000 dimensional statevector [4–6]. Quite apart from the challenges presented by preparation and measurement of quantum states, tackling the state reconstruction problem in the face of such complexity calls for sophisticated data processing techniques, which are the focus of this paper.

Tomography experiments involve subjecting a system, described by some unknown quantum state, to a well-defined measurement procedure and recording the measurement outcome. The central tenets of quantum theory place severe restrictions on one’s ability to charac-

terise an unknown quantum system given measurements made on only a single copy. One assumes, therefore, access to a large but finite number of copies of a system, all prepared in an identical quantum state. As the complexity of the quantum state grows, the number of detector counts for each state parameter necessarily shrinks. Particularly in optical systems, the prevalence of low detection efficiency exacerbates this problem, leaving the tomographer with a noisy data-set from which to make her inferences. In an idealised situation where all noise (including statistical) is absent, the true state  $\rho_{\text{true}}$  can be found exactly. Here we concentrate on the more realistic situation, and assume only that the measurement procedure is informationally complete (that is to say, the measurement record contains a nonzero amount of information available about each quantum state parameter), and turn our attention to the question of processing this data toward the most likely estimate of the unknown state.

In most cases, quantum tomographers must employ numerical techniques to search for the best estimate possible, given the data that has been collected. In this work, we analyse the algorithmic method of projected gradient descent as applied to quantum tomography, and benchmark it against a number of existing methods. The state-of-the-art with regards to full quantum tomography methods include the diluted iterative algorithm (DIA) [7, 8] and convex programming [9]. Both methods benefit from a theoretical guarantee: that they converge to the maximum likelihood (ML) state  $\rho_{\text{ML}}$  (discussed below). However, the DIA has been observed to converge slowly [10, 11], and convex programming solvers such as SDPT3 and SeDuMi are known to require computational time that scales poorly with non-sparse matrix dimensionality [12, 13].

A non-iterative quantum tomography method was devised by Smolin *et al.*, who showed that, if the mea-

surement operators are traceless and the noise is of the Gaussian type, the constrained maximum likelihood state  $\rho_{ML}$  can be retrieved in a single projection step from the unconstrained maximum likelihood state  $\chi_{ML}$ , which can be found very quickly using linear inversion [14]. With an implementation of this method's core algorithm using a GPU, Guo *et al.* were able to recover a simulated 14-qubit density matrix [15]. However, the restrictive above conditions motivate the search for more broadly applicable efficient techniques.

Projected gradient descent (PGD) methods are emerging as promising candidates for quantum tomography [10, 16]. We present three PGD algorithms that converge towards the maximum likelihood quantum state: projected gradient descent with backtracking (PGDB) [10, 16], Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [17] and projected gradient descent with momentum (PGDM). We provide evidence that they converge faster than both DIA and SDPT3, and scale more favourably than SDPT3. We also find that the PGD methods are very versatile in that one can model a wide variety of types of noise; in particular, the case of ill-conditioned measurements.

Gonçalves *et al.* [10] and Shang *et al.* [16] have both recently discussed PGDB as an efficient technique for quantum tomography: however, the algorithmic variants we introduce here can significantly outperform it. Furthermore, Shang *et al.* considered Pauli measurements, for which the technique from Ref. [14] is highly efficient. It is therefore vital to study the performance of projected gradient descent outside of this realm to establish its true usefulness. We here confirm that PGD methods continue to exhibit excellent properties in scenarios where the technique from Ref. [14] is not applicable.

The remainder of this paper is structured as follows: After giving an introduction to quantum tomography and the general idea of projected gradient descent, we lay out three PGD algorithms and discuss their performance: namely, convergence profiles and running time. To the best of our knowledge, FISTA has never previously been applied to quantum tomography and PGDM is a novel algorithm altogether. DIA and SDPT3, considered as current state-of-the-art algorithms, will serve as benchmarks by which the PGD approach will be judged. Finally, we report on the results of state reconstruction using pseudo-experimental data, i.e. simulations of realistic quantum tomographic experiments with noise. We consider three figures of merit for quantum tomographic techniques: the convergence time and the fidelity between the recovered state and the actual one  $\rho_{true}$ . Over a broad range of Hilbert space dimensions, we run the algorithms multiple times, each time with a randomly generated density matrices with fixed purity [18], and record the running times and fidelities.

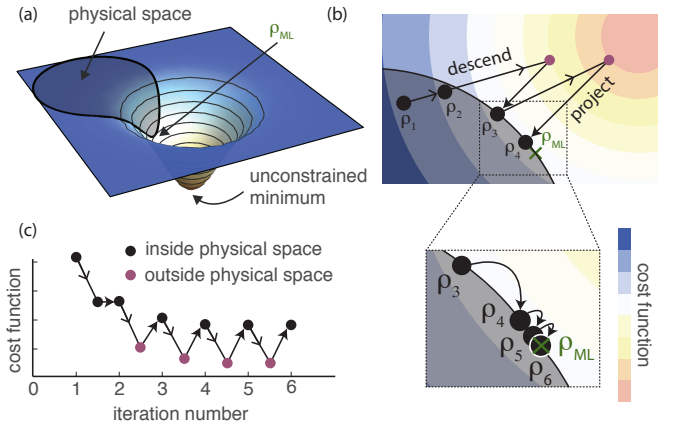


FIG. 1: a) Schematic showing the physical space as a convex subset of the set of unconstrained matrices. The minimum of the cost function is often outside of the physical space. b) Illustration of the PGD process for a qubit. A step in the gradient direction can yield a non physically-allowed density matrix, but the projection brings the estimate back into the desired search space, (i.e. the Bloch sphere in the case of a qubit). c) This process lowers the cost function, and is repeated until progress is sufficiently small and the final density matrix estimate is as close as desired to the maximum likelihood state  $\rho_{ML}$ .

## Quantum tomography

The Born rule,  $p_i = \text{Tr}(\Pi_i \rho)$ , being the central equation of quantum theory, encodes the probability  $p_i$  to obtain a certain measurement outcome given a particular quantum state. It involves a quantum state  $\rho$ , which takes the form of a  $d \times d$  positive-semidefinite matrix of unit trace, and an Hermitian measurement operator  $\Pi_i$ . Quantum tomography is essentially the process of finding the density matrix whose calculated outcome probabilities (for an informationally complete set of  $N$  operators) most closely match the experimentally observed data.

The probabilities  $p_i$  are of course not directly observable, only the number of clicks  $n_i$  recorded in a real measurement device after a finite number of trials. In the absence of noise, the probabilities relate to the number of clicks through a multiplicative factor  $r$ :  $n_i = r p_i$ . In real situations, there is a discrepancy between  $r p_i$  and  $n_i$  due to i) technical noise in the measurement device and ii) statistical uncertainty. Furthermore, if the noise is particularly severe, the matrix reconstructed with naive methods (such as linear inversion) will fail to qualify as a physical quantum state: the positivity or unit-trace properties can be violated. Multiple techniques have therefore been developed that allow one to search for an estimate matrix that is guaranteed to be physical. Examples include searching for the Cholesky factor  $T$  (where  $\rho = T T^\dagger$  is guaranteed positive) and using a Lagrange multiplier (to ensure unit trace) [19]. However, searching in the factored space can often lead to an ill-conditioned problem and slow convergence [16]. As we evidence, there are

advantages to be had by instead allowing the search to temporarily wander into unphysical territory.

The measurement operators, the expectation values  $p_i$  and the detector clicks  $n_i$  can be stacked into a matrix and two vectors, respectively:

$$A = \begin{pmatrix} \text{vec}(\hat{\Pi}_1)^T \\ \vdots \\ \text{vec}(\hat{\Pi}_N)^T \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_N \end{pmatrix}, \quad \text{and} \quad \mathbf{n} = \begin{pmatrix} n_1 \\ \vdots \\ n_N \end{pmatrix}, \quad (1)$$

where  $N$  is the total number of projectors. With the above notation, the expectation value vector reads  $\mathbf{p} = A \text{vec}(\rho)$ . The computation of this vector takes  $O(Nd^2)$  floating-point operations in general, but a lower computational complexity can be achieved when the operators originate from tensor products [16]. The accuracy of the maximum likelihood state depends significantly on the condition number (which is the ratio of maximum singular value to minimum singular value) of the measurement matrix  $A$  [20]. High condition numbers, which correspond to ill-conditioned measurement matrices, arise in the fields of detector tomography [21, 22] and superconducting artificial atoms [23].

The multiplicative factor  $r$  can readily be estimated if at least a subset  $\mathcal{Z}$  of the measurement matrix  $A$  forms a POVM, in which case the sum of the probabilities belonging to the set  $\mathcal{Z}$ ,  $\sum p_{\mathcal{Z}}$ , is independent of the state  $\rho$ . For example, if the measurement operators in  $\mathcal{Z}$  form a basis, the sum in question amounts to unity. For an arbitrary POVM, the best estimate for the multiplication factor is given by

$$r = \frac{d}{N_{\mathcal{Z}}} \sum_{j \in \mathcal{Z}} n_j, \quad (2)$$

where  $N_{\mathcal{Z}}$  is the number of projectors in the POVM. Moreover, the average number of clicks per outcome is  $r/d$ , and the total number of clicks for this POVM is  $rN_{\mathcal{Z}}/d$ .

#### Summary of PGD algorithms

The distance between  $p_i$  and  $n_i/r$  is to be considered as a ‘cost function’  $\mathcal{C}(\rho)$  in the sense of numerical optimisation. In the minimisation of a cost function, PGD algorithms are useful when one seeks a solution in a proper subset of a larger search space. Since physical quantum states, represented as unit-trace positive-semidefinite matrices, exist in a (convex) subset of the (convex) set of  $d \times d$  matrices, quantum tomography is indeed a problem of this kind. Projected gradient descent is an iterative procedure with two substeps. Starting with a well-chosen physical state, first a step is taken in the downhill direction of the cost function, which has the chance to result in a nonphysical matrix. Second, to

bring the estimate back within the constrained, physical space, we project it to the closest point in the solution space (for example, using a matrix norm). This two step process is then repeated until the cost function converges towards a low enough value. Since we are searching over a convex set, as long as the cost function is a strictly convex function of  $\rho$ , there will be a unique solution that minimises it. Fig. 1 shows the evolution of the density matrix estimate of a qubit through six iterations of PGD.

#### Maximum likelihood

We have yet to specifically define the figure of merit for *closeness* between the estimated probabilities and the outcome frequencies. Maximum likelihood analysis provides a principled way to derive such a figure of merit. When doing statistical estimation, it is necessary to operate within a statistical model or belief system: one approach is Bayesian estimation, which works with iteratively updating such beliefs using Bayes’ rule. Here, however, the beliefs are encoded in the likelihood function for a multinomial experiment:

$$\mathcal{L}(\rho) \propto \prod_i p_i^{n_i}. \quad (3)$$

Maximizing this function – i.e. finding the quantum state  $\rho$  that makes the observed data the most likely – is the most widely applied philosophy for tomography [7, 19, 24, 25]. Since the maximum-likelihood state  $\rho_{\text{ML}} = \max_{\rho} \mathcal{L}(\rho)$  is also the minimum of  $-\log \mathcal{L}$ , we can proceed by minimising the second function, and we may ignore any scale or shift by a constant that is independent of  $\rho$ . We therefore define the cost function

$$\mathcal{C}(\rho) = -\log \mathcal{L}(\rho) \quad (4)$$

that we seek to minimise. When the number of trials is large, this is well approximated by the Poisson-approximated Gaussian likelihood function  $\mathcal{C}(\rho) \approx -\log \mathcal{L}_P(\rho) = \boldsymbol{\nu}^T \boldsymbol{\nu}$  with  $\nu_i = (rp_i - n_i)/\sqrt{n_i}$ . Assuming Poisson-distributed data, the variance for outcome  $i$  is the number of clicks  $n_i$ . Hence  $\nu_i$  corresponds to the ratio of the error to the expected error on outcome  $i$ . The true density matrix gives an expected negative log-likelihood per outcome  $\mathcal{C}/N$  of unity because of the noise on the outcomes  $n_i$ . A value greater than unity indicates a poor density matrix estimate or an incomplete noise model, whereas a value smaller than unity is a sign of overfitting to noise. In general, the maximum likelihood density matrix overfits the data slightly [26], but one cannot achieve a better estimate in the absence of prior knowledge.

We are now ready to detail the algorithms for reconstructing the density matrix. In all of the following algorithms, the completely mixed state  $\rho_0 = I/d$  will serve

as the starting point. Our selection of four iterative algorithms are then defined by a recursion relation relating the density matrix at the next iteration to the matrix at the current iteration.

## RESULTS

### Projected Gradient Descent Algorithms

The process of any PGD algorithm involves steps in the general gradient direction, interspersed with leaps back into the constrained set [10, 16]. The simplest such algorithm can be written in a single line [4]:

$$\rho_{k+1} = \mathcal{P}[\rho_k - \delta \nabla \mathcal{C}(\rho_k)], \quad (5)$$

where  $\delta$  is a step size and  $\mathcal{P}[\cdot]$  is a projection onto the set of unit-trace positive matrices, seeking the ‘closest’ unit-trace positive matrix to its argument. Various approaches can be used to establish what ‘closest’ means (including operator norms, see Supplementary Information). We adopt the *simplex projection*  $\mathcal{P}[\cdot] \rightarrow \mathcal{S}[\cdot]$ , which essentially consists of transforming the eigenvalues of the density matrix such that the sum is unity trace and they are all positive [27]. If the multiplicative factor  $r$  is known or computed in advance, the version described in detail in Ref. [10] applies, otherwise the projection must instead be performed over the space of positive matrices, preserving the trace of the argument [14].

We now proceed to discuss three PGD algorithms which are all extensions of Eq. (5).

#### *Projected Gradient Descent with Momentum*

Here we augment the basic PGD algorithm of Eq. (5) with a technique borrowed the momentum-aided gradient descent method from the field of machine learning [28]. This technique stores a running weighted-average  $M_k$  of the log-likelihood gradient. This running average provides a memory of previous descent directions which is used to better estimate the next descent step. The core of this algorithm reads

$$\begin{aligned} M_{k+1} &= \zeta_k M_k - \gamma_k \nabla \mathcal{C}(\rho_k), \\ \rho_{k+1} &= \mathcal{S}(\rho_k + M_k), \end{aligned} \quad (6)$$

where  $\zeta_k$  codes for the level of ‘inertia’ for the descent direction, and  $\gamma_k$  is the step size. In general, these meta-parameters depend on the iteration number  $k$ , but can also be set as constants throughout the algorithm. We provide full pseudo-code for this and the other algorithms in the Methods section, as well as Matlab implementations.

#### *Fast Iterative Shrinkage-Thresholding Algorithm*

FISTA was first developed in the context of image denoising [17], but here we introduce the method for use in quantum state tomography with adapted refinements. In this implementation of PGD, the change in the iterate  $\rho_k$  is not always in the descent direction, i.e. the log-likelihood function can go up, but as we shall see it descends much faster on average than the basic gradient descent algorithm from Eq. (5). The core of the algorithm is given by

$$\rho_{k+1} = \mathcal{S} \left[ \rho_k + \frac{k-2}{k+1} (\rho_k - \rho_{k-1}) - \delta \nabla \mathcal{C}(\rho_k) \right], \quad (7)$$

where  $\delta$  is a step size.

#### *Projected gradient descent with backtracking*

This PGD method has recently been applied to quantum tomography simulations in Ref. [10]. A similar variant has been studied in Ref. [16], where the authors report on a hybrid method between the DIA and PGD. The PGDB algorithm, whose main characteristic consists of trying to find the maximum step size that reduces the negative log-likelihood, can be written as

$$\rho_{k+1} = (1 - \alpha) \rho_k + \alpha \mathcal{S}[\rho_k - \nabla \mathcal{C}(\rho_k)], \quad (8)$$

where  $\alpha$  is a parameter to be loosely optimised at each step through backtracking. Each iteration of this algorithm is guaranteed to lower the negative log-likelihood unless a stationary point is reached, in which case the stopping criterion is satisfied.

### Benchmark algorithms

#### *Diluted Iterative Algorithm*

The diluted iterative algorithm (DIA) is based on the gradient of the log-likelihood function. The algorithm can be simply stated with the following two iterative equations [7, 8]

$$\begin{aligned} R_k &= -H^{-1/2} [\nabla \mathcal{C}] H^{-1/2}, \\ \rho_k &= \frac{(I + \epsilon R_k) \rho_{k-1} (I + \epsilon R_k)}{\text{Tr}[(I + \epsilon R_k) \rho_{k-1} (I + \epsilon R_k)]}, \end{aligned} \quad (9)$$

where  $H = \sum_i \Pi_i / \sum_i p_i$ . The variable  $\epsilon$  is optimised at every iteration, such that it minimises the log-likelihood function, and can be implemented in various ways [8, 29]. The matrix  $H$  reduces to the identity (up to a constant) when all the measurement operators form a POVM. The DIA leaves the density matrix estimate  $\rho_k$  positive at every iteration.

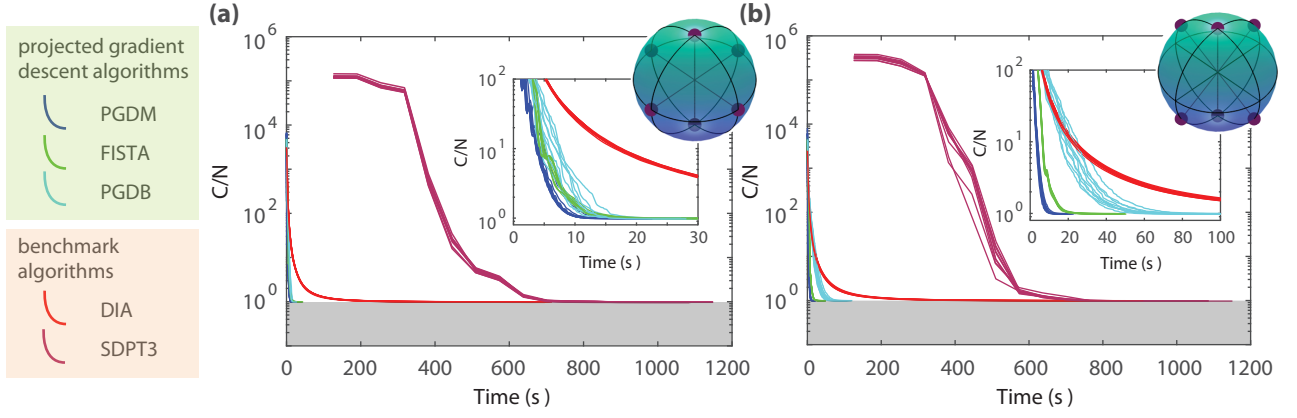


FIG. 2: Typical curves of the cost function versus running time. These simulations are performed on six-qubit systems using (a) Pauli measurements as indicated by the vectors on the Bloch sphere and (b) an ill-conditioned measurement matrix. The global minimum of the negative log-likelihood is expected to be at  $C/N \approx 1$ , around the top of the grey regions. Only for PGDM and FISTA can the cost function go up as a function of iteration number before reaching the ML state. The total running time for PGDB correlates highly with the measurement matrix condition number.

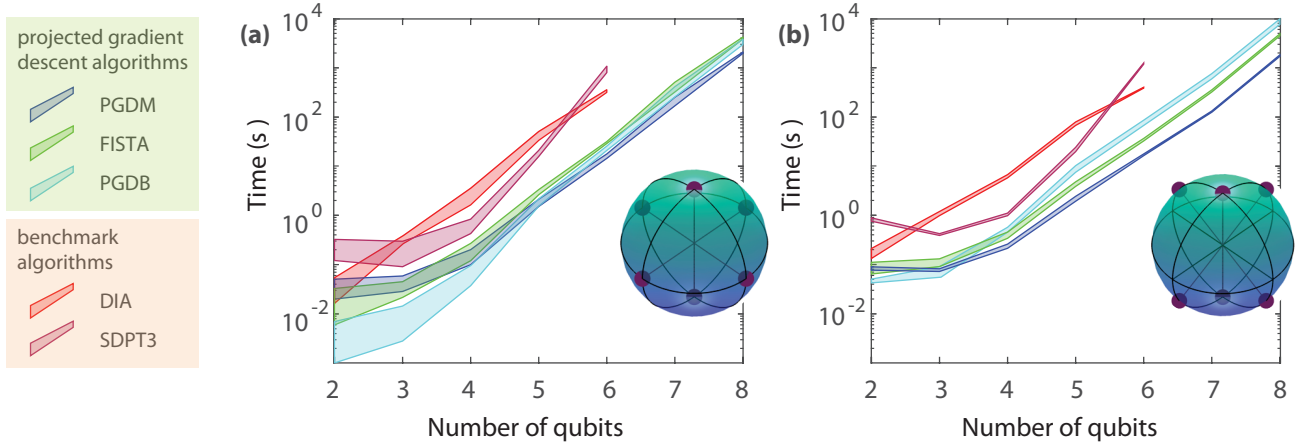


FIG. 3: Average running time to reach (sufficiently close to)  $\rho_{ML}$  for various system sizes. The measurement matrix is made of (a) Pauli measurements and (b) bases relatively close to each other, as indicated on the inset. The colored areas are bounded by one standard deviation above and below the mean running time. The gradient-based techniques have a computational complexity of  $O(Nd^2)$  while SDPT3 converges in time  $O(N^2d^2)$ .

It has been observed that the DIA converges quickly for the first few iterations and converges very slowly later [10, 11, 16, 21]. In the Results section, we corroborate these observations.

### *Semidefinite programming*

As we emphasised above, quantum tomography is often equivalent to minimising a convex function over a convex set. In the field of numerical optimisation, a problem is considered effectively solved if it can be cast into this form, partly because of the powerful and efficient algorithms and software packages that are available, and also because of the guarantee of global optimality for the solutions that they find. Such a software package therefore makes a natural benchmark for quantum tomography al-

gorithms, with the understanding that (because of their general purpose nature) they are not optimised for full tomography and unlikely to be as fast as other methods. We use the CVX software environment; and the SDPT3 solver, which is an example of an infeasible path-following algorithm.

### **Simulation**

We perform quantum tomography simulations on multi-qubit systems with all of the techniques mentioned in the previous section. When using canonical Pauli measurements, all of the techniques are found to work well in that they all recover  $\rho_{ML}$ . Since the simulations consistently reach high likelihoods in practice, we concentrate on the total computation time. The exit criterion for all

techniques – except for SDPT3 whose code we do not change – is such that when the average gradients of the last 20 iterations is sufficiently small, the optimisation procedures are terminated.

Examples of convergence curves for 6-qubit systems using Pauli measurements are shown in Fig. 2. The details of the implementations and simulated data are provided in the Methods section. As already remarked in Refs. [10, 11, 16, 21], the DIA displays fast convergence in the first few iterations but requires many more to finally satisfy the exit criterion. SDPT3 converges in between only 10 and 15 iterations, but each iteration has a computational complexity of  $O(N^2 d^2)$ , rendering it slow in high dimensions.

We put the tomographic techniques to the test using ill-conditioned measurement matrices [20, 21, 23], see Fig. 2 b). If the measurement operators are limited to a restricted region of the Hilbert space, the condition number of the measurement matrix is greater than unity, and the error on the final density matrix estimate will necessarily increase [20]. Here, the measurement matrix is built using the bases

$$\begin{aligned} &[0 \ 1]; [1 \ 0]; [\cos(\beta/2), \sin(\beta/2)]; [\sin(\beta/2), \cos(\beta/2)]; \\ &[\cos(\beta/2), i\sin(\beta/2)]; [\sin(\beta/2), i\cos(\beta/2)] \end{aligned} \quad (10)$$

with  $\beta = \pi/4$  for regular canonical Pauli operators and  $\beta = \pi/3$  for the ill-conditioned case; see the Bloch spheres on Fig. 2 for an illustration of these vectors.

Gonçalves *et al.* provide a proof of the monotonicity of  $\mathcal{C}$  for PGDB, that is to say that the cost function never increases in this algorithm. By contrast, PGDM and FISTA are both algorithms for which the cost function may increase from one iteration to the next, but interestingly, this tends to speed up their performance relative to PGDB with regards to the ill-conditioned measurement matrices.

We show the running time of each algorithm as a function of Hilbert space dimensionality (number of qubits) in Fig. 3. The speed-up of PGDM over PGDB for high-dimensions is present in both panels of Fig. 3, but particularly pronounced for the quantum state reconstruction task based on the ill-conditioned measurement matrices, where PGDM is about ten times faster than PGDB on average for the eight and seven-qubit cases.

The running times of PGDM and FISTA algorithms are more resilient to a condition number change than PGDB, due to the fact that the number of PGDM and FISTA iterations required to reach  $\rho_{\text{ML}}$  grows very little as a function of the measurement matrix condition number. Fig. 4 a) illustrates this dependence. The semidefinite programming technique does not depend on the condition number: in our simulations, SDPT3 always took about 15 iterations to reach  $\rho_{\text{ML}}$ .

There exists a monotonic relationship between ill-posedness and the accuracy of the recovered state: for

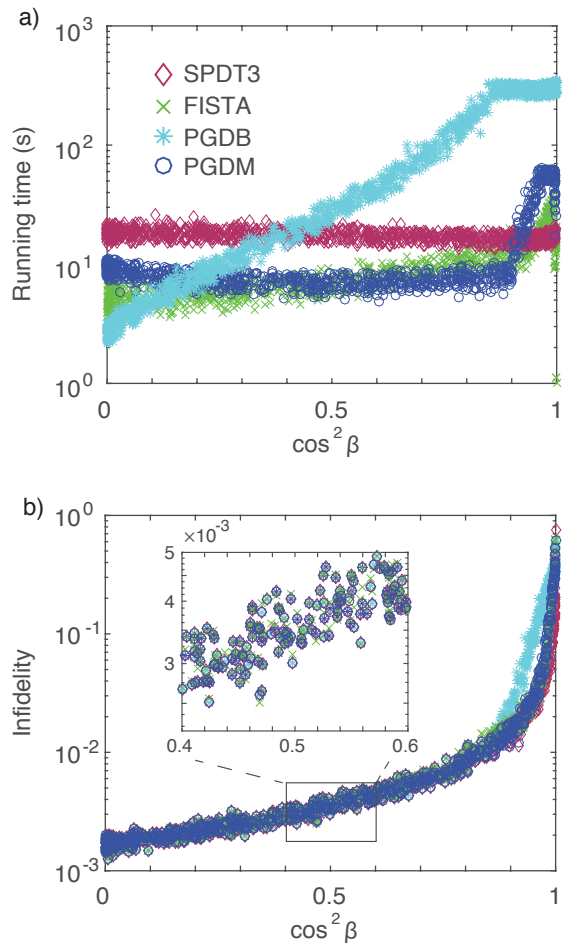


FIG. 4: Scatter plot of (a) the time performance and (b) the infidelity between the recovered five-qubit states and the true states as a function of the ill-posedness of the measurement matrix. The PGDB method saturates in (a) because it reaches the maximum number of iterations set in the program. The inset in (b) shows that the methods converge towards the same fidelity, indicating that they reach  $\rho_{\text{ML}}$ . The results of FISTA differ slightly because this method oscillates around  $\rho_{\text{ML}}$  for a large number of iterations. The number of events per outcome is set to  $10^4$ .

a fixed number of events per measurement, the more ill-posed the measurement matrix is, the lower the fidelity between the recovered state and the true one. This relationship is illustrated in Fig. 4b, where the vertical axis corresponds to one minus the fidelity  $f(\rho_1, \rho_2) = \text{Tr}(\sqrt{\sqrt{\rho_1}\rho_2\sqrt{\rho_1}})$  between the recovered density matrix and the true state. The extreme cases, i.e.  $\cos^2 \beta = 0$  and  $\cos^2 \beta = 1$ , correspond to mutually unbiased bases and a single basis measurement, respectively. We show a zoomed subset of this data around the intermediate angle, i.e.  $\cos^2 \pi/4 = 0.5$ , that gives statistical evidence that the projected gradient descent techniques consistently converge to  $\rho_{\text{ML}}$ .

## DISCUSSION

A key advantage of the PGD techniques is their versatility. They successfully and quickly converge to the maximum likelihood state in a wide range of cases, regardless of the desired accuracy and whether the measurement matrix is well- or ill-conditioned. PGDM and FISTA – our new PGD algorithms – are shown to be especially well suited for ill-conditioned problems.

Quantum tomography includes three main subfields: detector, process and state tomography. State tomography algorithms are straightforwardly transferable to process tomography, but applying the same algorithms to detector tomography with success is not trivial. The problem of detector tomography lies in characterising an unknown detector POVM from an informationally complete set of known states. If the tomographer probes an optical detector with coherent states, the problem of detector tomography is ill-conditioned, and like the density matrix, the POVM elements must be positive-semidefinite [22]. Currently, the state-of-the-art algorithms to solve this problem are semidefinite program solvers such as SeDuMi [21]. Because PGDM and FISTA perform well in the case of ill-conditioned measurement matrices, these algorithms hold great promise for optical detector tomography. One avenue for future work is the application of these two algorithms to the characterisation of detector POVMs in high dimensions.

In summary, the PGD techniques have proven their worth in that they all converge towards the maximum likelihood density matrix reliably. The different PGD techniques complement each other such that PGDB is fastest in low dimensions and, according to our simulations, PGDM is fastest beyond five-qubit systems. Further, we find that the PGD techniques reach  $\rho_{\text{ML}}$  significantly faster than the DIA and SDPT3 in the vast majority of scenarios, thus surpassing the state-of-the-art techniques with regards to assumption-free quantum state tomography.

## Methods

### Rank-1 projectors

To reduce the memory requirements, we chose to use rank-1 measurement operators in the simulation. Instead of matrix operators, the measurements take the form of  $d$ -dimensional vectors. Given rank-1 projectors  $|\phi_i\rangle\langle\phi_i|$ , the Born rule is written  $p_i = \langle\phi_i|\rho|\phi_i\rangle$  and the measurement matrix takes the following form

$$\mathcal{A} = \begin{pmatrix} \langle\phi_1| \\ \vdots \\ \langle\phi_N| \end{pmatrix}. \quad (11)$$

In this compact notation, the measurement matrix is  $(N \times d)$ -dimensional, thus requiring  $d$  times less RAM memory compared to the full-rank case.

The gradient of the Gaussian negative log-likelihood function is compactly written as

$$\nabla \log \mathcal{L}_G(\rho) = 2G^\dagger \mathcal{A}^*, \quad (12)$$

where the elements of the  $G$  matrix are defined:  $G_{i,j} = \mathcal{A}_{i,j}(rp_i - n_i)$ . The computation of the above gradient takes  $O(Nd^2)$  floating-point operations. This is the gradient that we use in the PGD algorithms for the simulations.

For all simulations in the main text, the average number of events per outcome  $r/d$  is set to  $10^4$ . Density matrices were chosen randomly in the Haar sense but always with a purity of 0.5. All simulations were performed on a single thread on an Intel Xeon Haswell processor.

### PGD algorithms

The pseudo code for PGDM, FISTA and PGDB is given in Algorithms 1, 2 and 3. The symbol  $\circ$  corresponds to the Hadamard product (or element-wise multiplication).

---

#### Algorithm 1 PGDM

---

```

1:  $k = 0$ 
2: Initial estimate and momentum matrix:  $\rho_0 = I$ ,  $M_0 = 0$ 
3:  $currentMagnitude = \lceil \log_{10} \mathcal{C}_P(\rho_0) \rceil$ 
4: Set step size and inertia:  $\gamma = (2rd)^{-1}$ ,  $\zeta = 0.95$ 
5: while  $\sum_{j=1}^{20} |\mathcal{C}_P(\rho_j) - \mathcal{C}_P(\rho_{j-1})| > 10^{-5}$  do
6:   Projection:  $\rho_k = \mathcal{S}(\rho_k)$ 
7:   Estimate probabilities:  $\mathbf{p}_k = \sum_j [\mathcal{A} \circ (\mathcal{A}^* \rho_k)]_{i,j}$ 
8:   Calculate log-likelihood:  $\mathcal{C}_P(\rho_k) = \boldsymbol{\nu}^T \mathbf{p}_k / N$ 
9:   Compute gradient:  $\nabla \mathcal{C}_G(\rho_k) = 2G^\dagger \mathcal{A}^*$ 
10:   $currentMagnitude = \lceil \log_{10} \mathcal{C}_P(\rho_k) \rceil$ 
11:  if  $currentMagnitude < previousMagnitude$  then
12:    Update inertia:  $\zeta_k = (1 - (1 - \zeta_k) * 0.95)$ 
13:     $previousMagnitude = currentMagnitude$ 
14:    Update momentum:  $M_{k+1} = \zeta_k M_k - \gamma \nabla \mathcal{C}_G(\rho_k)$ 
15:    Update density matrix:  $\rho_{k+1} = \rho_k + M_{k+1}$ 
16:     $k = k + 1$ 
17: Final projection:  $\rho_{\text{final}} = \mathcal{S}(\rho_{k+1})$ 
18: Return  $\rho_{\text{final}}$ 

```

---



---

\* email: J.Leach@hw.ac.uk

- [1] W.-B. Gao, C.-Y. Lu, X.-C. Yao, P. Xu, O. Gühne, A. Goebel, Y.-A. Chen, C.-Z. Peng, Z.-B. Chen, and J.-W. Pan, Nature Phys. **6**, 331 (2010).
- [2] P. Schindler, J. T. Barreiro, T. Monz, V. Nebendahl, D. Nigg, M. Chwalla, M. Hennrich, and R. Blatt, Science **332**, 1059 (2011).



---

**Algorithm 2 FISTA**


---

```

1:  $k = 0$ 
2: Initial estimate and momentum matrix:  $\rho_0 = I, M_0 = 0$ 
3: Set step size:  $\delta = (10d)^{-1}$ 
4: while  $\sum_{j=1}^{20} |\mathcal{C}_P(\rho_j) - \mathcal{C}_P(\rho_{j-1})| > 10^{-6}$  do
5:   Projection:  $\rho_k = \mathcal{S}(\rho_k)$ 
6:   Estimate probabilities:  $\mathbf{p}_k = \sum_j [\mathcal{A} \circ (\mathcal{A}^* \rho_k)]_{i,j}$ 
7:   Calculate log-likelihood:  $\mathcal{C}_P(\rho_k) = \boldsymbol{\nu}^T \boldsymbol{\nu} / N$ 
8:   Compute gradient:  $\nabla \mathcal{C}_G(\rho_k) = 2G^\dagger \mathcal{A}^*$ 
9:    $\rho_{k+1} = \rho_k + (k-2)(\rho_k - \rho_{k-1})(k+1)^{-1} - \delta \nabla \mathcal{C}_G(\rho_k)$ 
10:   $k = k + 1$ 
11: Final projection:  $\rho_{\text{final}} = \mathcal{S}(\rho_{k+1})$ 
12: Return  $\rho_{\text{final}}$ 

```

---



---

**Algorithm 3 PGDB**


---

```

1:  $k = 0$ 
2: Initial estimate and momentum matrix:  $\rho_0 = I, M_0 = 0$ 
3: Set metaparameters:  $\mu = 1, \ell = 10^{-4}$ 
4: while  $\sum_{j=1}^{20} |\mathcal{C}(\rho_j) - \mathcal{C}(\rho_{j-1})| > 10^{-5}$  do
5:   Projection:  $\rho_k = \mathcal{S}(\rho_k)$ 
6:   Estimate probabilities:  $\mathbf{p}_k = \sum_j [\mathcal{A} \circ (\mathcal{A}^* \rho_k)]_{i,j}$ 
7:   Calculate log-likelihood:  $\mathcal{C}(\rho_k) = \boldsymbol{\nu}^T \boldsymbol{\nu} / N$ 
8:   Compute gradient:  $\nabla \mathcal{C}(\rho_k) = 2P^\dagger \mathcal{A}^*$ 
9:    $\rho'_k = \mathcal{S}(\rho_k - \mu^{-1} \nabla \mathcal{C}(\rho_k))$ 
10:   $D = \rho'_k - \rho_k$ 
11:  Line search initialisation:  $\alpha_k = 1$ 
12:   $\mathcal{C}'_G(\rho_k) = \mathcal{C}_G(\rho_k) + \ell \alpha_k \text{Tr}[D \nabla \mathcal{C}_G(\rho_k)]$ 
13:  while  $\mathcal{C}_G(\rho_k + \alpha_k D) > \mathcal{C}'_G(\rho_k)$  do
14:    Line search:  $\alpha_k = \alpha_k / 2$ 
15:     $\mathcal{C}'_G(\rho_k) = \mathcal{C}_G(\rho_k) + \ell \alpha_k \text{Tr}[D \nabla \mathcal{C}_G(\rho_k)]$ 
16:  Update density matrix:  $\rho_{k+1} = \rho_k + \alpha_k D_k$ 
17:   $k = k + 1$ 
18: Final projection:  $\rho_{\text{final}} = \mathcal{S}(\rho_{k+1})$ 
19: Return  $\rho_{\text{final}}$ 

```

---

- [3] X.-C. Yao, T.-X. Wang, P. Xu, H. Lu, G.-S. Pan, X.-H. Bao, C.-Z. Peng, C.-Y. Lu, Y.-A. Chen, and J.-W. Pan, *Nature Comm.* **6**, 225 (2012).
- [4] E. Bolduc, G. Gariépy, and J. Leach, *Nature Comm.* **7** (2016).
- [5] M. Malik, M. Mirhosseini, M. Lavery, J. Leach, M. J. Padgett, and R. W. Boyd, *Nat Commun* **5**, 3115 (2014).
- [6] X. C. Yao, T. X. Wang, P. Xu, H. Lu, G. S. Pan, X. H. Bao, C.-Z. Peng, C.-Y. Lu, Y.-A. Chen, and J.-W. Pan, *Nature Comm.* **6**, 225 (2012).
- [7] J. Řeháček, Z. Hradil, and M. Ježek, *Phys Rev A* **63**, 040303 (2001).
- [8] J. Řeháček, Z. Hradil, E. Knill, and A. I. Lvovsky, *Phys Rev A* **75**, 042108 (2007).
- [9] M. Grant, S. Boyd, and Y. Ye (2008).
- [10] D. S. Gonçalves, M. A. Gomes-Ruggiero, and C. Lavor, *Optimization Methods and Software* **31**, 328 (2016).
- [11] G. B. Silva, S. Glancy, and H. M. Vasconcelos, *arXiv preprint arXiv:1604.00321* (2016).

- [12] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, *Optimization Methods and Software* **11**, 545 (1999).
- [13] J. F. Sturm, *Optimization Methods and Software* **11**, 625 (1999).
- [14] J. A. Smolin, J. M. Gambetta, and G. Smith, *Phys Rev Lett* **108**, 070502 (2012).
- [15] Z. H. Guo, H.-S. Zhong, Y. Tian, D. Dong, B. Qi, L. Li, Y. Wang, F. Nori, G.-Y. Xiang, C.-F. Li, et al., *New J. Phys.* **18**, 083036 (2016).
- [16] J. Shang, Z. Zhang, and H. K. Ng, *arXiv preprint arXiv:1609.07881* (2016).
- [17] A. Beck and M. Teboulle, *SIAM Journal on Imaging Sciences* **2**, 183 (2009).
- [18] K. Życzkowski, K. A. Penson, I. Nechita, and B. Collins, *Journal of Mathematical Physics* **52**, 062201 (2011).
- [19] K. Banaszek, G. M. D'Ariano, M. G. A. Paris, and M. F. Sacchi, *Phys Rev A* **61**, 10304 (2000).
- [20] A. Miranowicz, K. Bartkiewicz, J. Peřina Jr, M. Koashi, N. Imoto, and F. Nori, *Phys Rev A* **90**, 062123 (2014).
- [21] A. Feito, J. S. Lundeen, H. Coldenstrodt-Ronge, J. Eisert, M. B. Plenio, and I. A. Walmsley, *New J. Phys.* **11**, 093038 (2009).
- [22] J. S. Lundeen, A. Feito, H. Coldenstrodt-Ronge, K. L. Pregnell, C. Silberhorn, T. C. Ralph, J. Eisert, M. B. Plenio, and I. A. Walmsley, *Nature Phys.* **5**, 27 (2009).
- [23] R. Bianchetti, S. Filipp, M. Baur, J. M. Fink, C. Lang, L. Steffen, M. Boissonneault, A. Blais, and A. Wallraff, *Phys Rev Lett* **105**, 223601 (2010).
- [24] M. S. Kaznady and D. F. James, *Phys Rev A* **79**, 022109 (2009).
- [25] D. F. James, P. G. Kwiat, W. J. Munro, and A. G. White, *Phys Rev A* **64**, 052312 (2001).
- [26] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C*, vol. 2 (Cambridge university press Cambridge, 1996).
- [27] C. Michelot, *Journal of Optimization Theory and Applications* **50**, 195 (1986).
- [28] I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton, *ICML (3)* **28**, 1139 (2013).
- [29] D. S. Gonçalves, M. A. Gomes-Ruggiero, and C. Lavor, *Quantum Information & Computation* **14**, 966 (2014).

### Acknowledgements

E.B. acknowledges the financial support of the FQRNT, grant #176729. J.L. acknowledges the financial support of the Engineering and Physical Sciences Research Council (EPSRC, UK, Grants EP/M006514/1 and EP/M01326X/1). G.C.K. was supported by the Royal Commission for the Exhibition of 1851. E.M.G. acknowledges support from the Royal Society of Edinburgh and the Scottish Government.